BMC Genomics

RESEARCH

Open Access

Systematic revelation and meditation on the significance of long exons using representative eukaryotic genomes

Yuange Duan[1] and Qi Cao[2*]

Abstract

Background Long exons/introns are not evenly distributed in the genome, but the biological significance of this phenomenon remains elusive.

Materials and methods Exon properties were analyzed in seven well-annotated reference genomes, including human and other representative model organisms: mouse, fruitfly, worm, mouse-ear cress, corn, and rice.

Results In all species, last exons in genes tend to be the longest. Additionally, we found that (1) canonical splicing motifs are strongly underrepresented in 3'UTR; (2) Last exons tend to have low GC content; (3) Comparing with other species, first exons in *D. melanogaster* genes demonstrate lower GC content than internal exons.

Conclusions It cannot be excluded that last exons of genes exert essential regulatory roles and is subjected to natural selection, exhibiting differential splicing tendency, and GC content compared to other parts of the gene body.

Keywords Reference genomes, Long exons, 3'UTR, Gene structure, Regulatory role

Background

Old trees with new blossom in this omics era

Genomic data are "gold mines" that contain numerous hidden information to be discovered [1–3]. Genomes can be studied either extensively, using generated new sequencing data [4–6], or intensively. A highly successful case is the reuse of transcriptome data to study the identification [7–9], regulation [10], evolution [11], and interpretation [12] of RNA editing. Many RNA-Seq data were originally generated for profiling gene expression levels, but researchers smartly take advantage of these data to unravel the biological significance of RNA editing. Another example is the reuse of population genomics data which were originally used for the inference of population history/differentiation. These data could be reused for analyzing the selective preference on synonymous mutations [13, 14] and RNA editing events [15, 16].

Initial discoveries on the distribution of long exons and introns

Reference genomes represent the most fundamental data for most bioinformatic analyses [1, 2, 3, 17, 18], but we are far from getting full understanding of the genomes. The length, structure, GC content, and conservation of the genome sequences are continuously being uncovered [19, 20]. A typically interesting case is the uneven distribution of long exons and introns in genomes [21, 22]. At the beginning of post-genomic era, by re-analyzing

*Correspondence:
Qi Cao
caoqi@bjmu.edu.cn
[1]Department of Entomology and MOA Key Lab of Pest Monitoring and Green Management, College of Plant Protection, China Agricultural University, Beijing 100193, China
[2]Health Science Center, International Cancer Institute, Peking University, Beijing 100191, China

the reference genomes of a handful of species, researchers found that extremely long exons or introns are rare [23] and that there is an ordinal reduction (from 5' to 3') of exon and intron lengths within a gene [24], where the first intron tends to be the longest [25]. However, despite the overall reduction of exon/intron length from 5' to 3', last exons turn out to be the longest [26]. Then, researchers tried to find out the reason for such unevenly distributed long exons and introns. The first intron is closest to the promoter of a gene and contains many regulatory elements to facilitate gene expression [27]. The longer last exon in human and mouse might be explained by the avoidance of introns in 3'UTR (untranslated region) due to the intron-dependent nonsense-mediated decay (NMD), but for fruitfly and *Arabidopsis* that lack this NMD mechanism, it remains unclear why last exons are still constrained. In addition to the discussion on the lengths of exons and introns, revisiting existing genomes also produces other aspects of new findings such as the trade-offs between synthetic cost and translation efficiency of transcripts [28] and the inference of evolutionary trajectory of particular orthologous genes and sites [29, 30].

Aims and scopes

Most of the previous studies on long exons and introns only focused on a single lineage like mammal or plant. We need to broaden the generality of this phenomenon and meditate on the underlying biological significance and evolutionary driving force. In this work, we retrieve seven well-annotated reference genomes: human (*Homo sapiens*), mouse (*Mus musculus*), fruitfly (*Drosophila melanogaster*), worm (*Caenorhabditis elegans*), mouse-ear cress (*Arabidopsis thaliana*), corn (*Zea mays*), and rice (*Oryza sativa*) that cover vertebrate, invertebrate, and plant. We first confirmed several hypotheses such as intron length gradually decreases from 5' to 3' and that last exons tend to be the longest. Then, we made the following new findings: (1) In all species, the present 3'UTR sequence significantly avoids splicing junctions or even canonical splicing motifs. (2) In all species, last exons tend to have the lowest GC content, which is possibly connected to the maintenance of m^6A modification and the accessibility of microRNAs to 3'UTR. (3) Comparing with other species, first exons in *D. melanogaster* genes demonstrate lower GC content than "internal" exons, presumably due to the selection force on translational initiation rate in 5'UTR and the preference on synonymous codon usage in coding sequence (CDS). In conclusion, last exons exhibit differential splicing tendency and GC content compared to other parts of the gene body.

Results

Basic statistics of reference genomes: genes, transcripts, exons, and introns

We downloaded seven well-annotated reference genomes: human (*Homo sapiens*), mouse (*Mus musculus*), fruitfly (*Drosophila melanogaster*), worm (*Caenorhabditis elegans*), mouse-ear cress (thale cress, *Arabidopsis thaliana*), corn (*Zea mays*), and rice (*Oryza sativa*). The choice of these species considers their representativeness among vertebrates, invertebrates, and plants, and also due to the fact that they are the most well-annotated model genomes currently. The detailed genome versions are provided in **Materials and Methods**. Here, we began with some basic statistics on the numbers of genes, transcripts, and exons (Table 1).

The number of genes per species ranges from 17,478 in *D. melanogaster* to 58,051 in *H. sapiens*, where the two mammalian species have more genes than other species. *D. melanogaster* has remarkably lower gene number than the second lowest species *A. thaliana* (Table 1). This suggests a tendency of complex gene-regulation networks in mammals. But we cannot rule out the explanation by Ohno's hypothesis (2R hypothesis) [31] that two or more rounds of whole genome duplications might have occurred at the common ancestor of vertebrates, leading to the diversification of genes. Moreover, the data suggest that mammals have larger numbers of transcripts (mRNAs, or splicing isoforms) per gene compared to invertebrates and plants (Table 1). It cannot be excluded that mammals might have greater demands on molecular diversity than others, or it might be otherwise explained by the greater efforts in annotating the human and mouse genomes. Moreover, a significant positive correlation is observed between the number of exons per gene and the

Table 1 Basic statistics of the reference genomes

Species	H. sap	M. mus	D. mel	C. ele	A. tha	Z. may	O. sat
# of genes	58,051	48,526	17,478	46,739	32,833	44,303	38,993
# of transcripts	198,002	117,762	34,472	58,941	54,013	77,341	45,973
Transcript/gene	3.41	2.43	1.97	1.26	1.65	1.75	1.18
# of exons	1,182,163	721,666	186,737	251,447	313,952	460,812	200,789
Genes > 20 exons	4.2%	4.7%	0.55%	0.55%	1.6%	1.4%	0.95%
Exon/transcript	5.97	6.13	5.42	4.27	5.81	5.96	4.37
Unique exon/gene	9.84	8.17	4.76	3.37	5.88	5.84	4.58
Transcripts/gene *versus* exons/gene	Pearson's $Cor = 0.93$, $P = 0.0024$						

number of transcripts per gene (Table 1 and Additional file 1: Supplementary Figure S1).

Interestingly, although on average each gene has more than 1.6 transcripts in many species, from the distribution we see that actually most genes only have one transcript (Fig. 1A, the leftmost bar is the highest). While we acknowledge that annotations might not include minor transcript forms even in model genomes, our results are based on the currently available genome data and we temporarily presume that they are accurate. It suggests that the average number of transcripts per gene is skewed by the outliers with an extremely large number of transcripts. The same goes for the number of exons per transcript: in all species, on average one transcript has more than four exons but in fact most transcripts only have one or two exons (Fig. 1A).

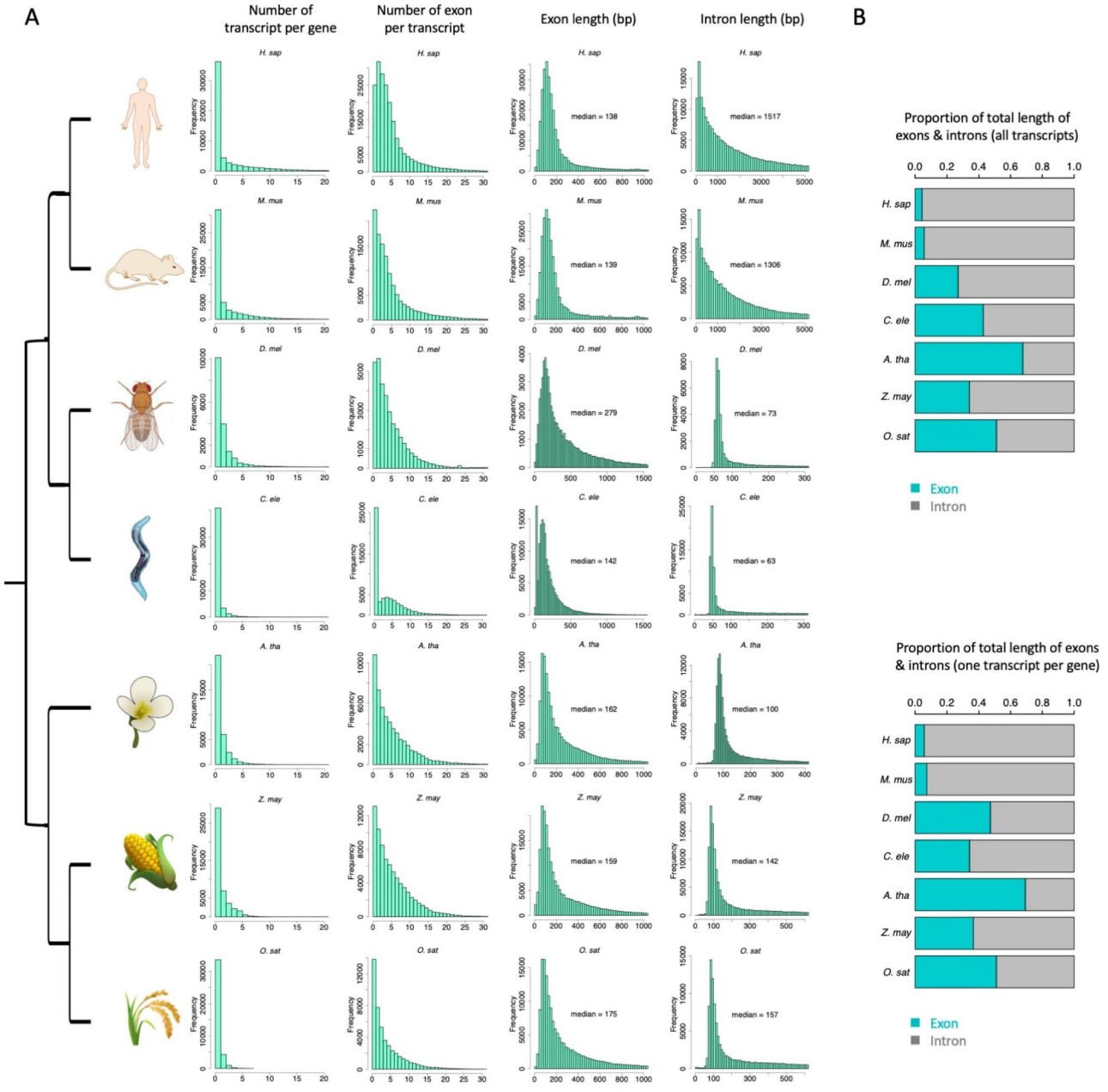

Fig. 1 Statistics on the numbers and lengths of transcripts, exons, and introns. The seven representative species are displayed in a phylogenetic order with unscaled branch length. (**A**) Histograms showing the distribution of the number of transcripts per gene, the number of exons per transcript, and the lengths of exons and introns. X-axis is the number or length, and Y-axis is the frequency, meaning how many elements have this number/length. The median lengths of exons and introns in each species are provided in the panel. (**B**) Barplots showing the proportions of total length of exons and introns. X-axis is the proportion. Bars show different species. The upper panel comes from all annotated transcripts. In the lower panel, only the transcript with the largest number of exons is selected for each gene

The same "outlier effect" is observed in the comparison of total lengths of exons and introns. By considering all the annotated transcripts in each species, we found that exons only make up 4.4% and 5.7% of gene regions in human and mouse, respectively (Fig. 1B). In contrast, in the fruitfly genome, a much higher fraction of exon regions is observed (27.0%), and in the thale cress genome, 67.8% of genic regions belong to exons, which is twice the total length of introns. This pattern holds true when we only consider one transcript per gene instead of all transcripts (Fig. 1B). The selected transcript is the one with the largest number of exons. Interestingly, in fruitfly, although the total length of introns is higher than that of exons, their length distribution shows that the median exon length (279 bp) is 3.8 times higher than the median intron length (73 bp) (Fig. 1A). This contradiction can be explained by the fact that a few extremely long introns drastically elevate the average intron length, leading to an overall larger fraction of intronic regions. But the few outliers have little effect on the median intron length since most introns are still shorter than exons. For example, the longest intron in fruitfly is the first intron of gene *Myo81F* (Additional file 1: Supplementary Figure S2, FBgn0267431: FBtr0392909) encoding Myosin 81 F with actin filament binding and microfilament motor activity (http://flybase.org/reports/FBgn0267431.htm). This intron is 268,107 bp in length, 3,673 times higher than the median intron length (73 bp). In fact, only 1,476 (2.98%) of the total introns are longer than 10,000 bp. These 2.98% introns have elevated the average intron length for 2.3 times (from 562.8 bp to 1275.5 bp), but only increase the median length for 2 bp (from 71 bp to 73 bp).

In human and mouse, although the median exon length is much lower than the median intron length (roughly a 1/10 ratio, Fig. 1A), this 1/10 foldchange is insufficient to explain that the total exons only make up 5% of the mammalian gene regions (Fig. 1B). This again implies the existence of "outlier introns" that remarkably skew the average intron length.

Our results suggest that the notion of "introns are much longer than exons" is only true in the two mammals (human and mouse) we tested. For fruitfly, the overall larger fraction of total intron length compared to total exon length is due to a few "outlier introns", and the majority of introns are actually shorter than exons. Moreover, in three plants, the median length of exons is higher than that of introns, showing an opposite trend to the traditional notion. We reiterate that these conclusions are made by trusting the accuracy and completeness of the current reference genomes, not considering potential minor transcript forms that are missed from the annotation.

The lengths of individual exons decrease with the number of exons per gene

In humans, one gene (transcript) can have as many as 363 exons (ENSG00000155657: ENST00000589042, protein: Titin, highly abundant in striated muscle, https://www.genecards.org/cgi-bin/carddisp.pl?gene=TTN_keywords=ENSG00000155657), and its ortholog in mouse has 347 exons (ENSMUSG00000051747: ENSMUST00000099981) (Additional file 1: Supplementary Figure S2). We wonder whether exon length differ between genes with numerous exons and those with very few exons. For each gene, we selected the transcript with the most exons and ranked different genes with increasing number of exons. Interestingly, in all seven species, the lengths of individual exons decrease with the number of exons per gene (Fig. 2), and this correlation is more evident in fruitfly and plants. Such finding is potentially explained by the intron-late hypothesis [32, 33] that introns insert into genes during evolution so that the original exon was split into multiple new exons: while increasing the number of exons, the length of each individual exon was reduced. The only exception of this trend is *C. elegans* where the single exon genes (mostly non-coding RNAs) tend to be shorter than others (Fig. 2).

Interestingly, in fruitfly, we observed that the individual intron length is significantly positively correlated with the number of introns per gene ($Rho = 0.97$, $P = 7.8E\text{-}6$), but this pattern is not observed in other species. In human, corn, and rice, this correlation is negative; and in mouse and thale cress, no significant correlation is observed (Fig. 2).

Note that in this part, only genes with no more than 20 exons (19 introns) are shown. In fact, in all species we used, only < 5% of the total genes have more than 20 exons (Table 1). Especially, fruitfly and worm only have 0.55% such genes. As a result, bins with rank > 20 will have variable length distributions that lack statistical power and skew the global trend. We therefore only show the genes with no more than 20 exons.

Intron length decreases from 5′ to 3′ while last exons are the longest

The above analysis distinguishes how many exons/introns a gene has but has pooled all exons/introns within each gene to see their length distribution. Next, for these multi-exon and multi-intron genes, we wonder whether individual exon/intron length correlates with the positional order (rank within gene) of the exon/intron.

As we have introduced, previous studies have investigated a few species and found that first introns and last exons tend to be the longest [24, 26]. To test the generality of this pattern, we retrieved the genes with no more than 20 exons (19 introns). We display the length of each exon/intron from 5′ to 3′. Figure 3 shows

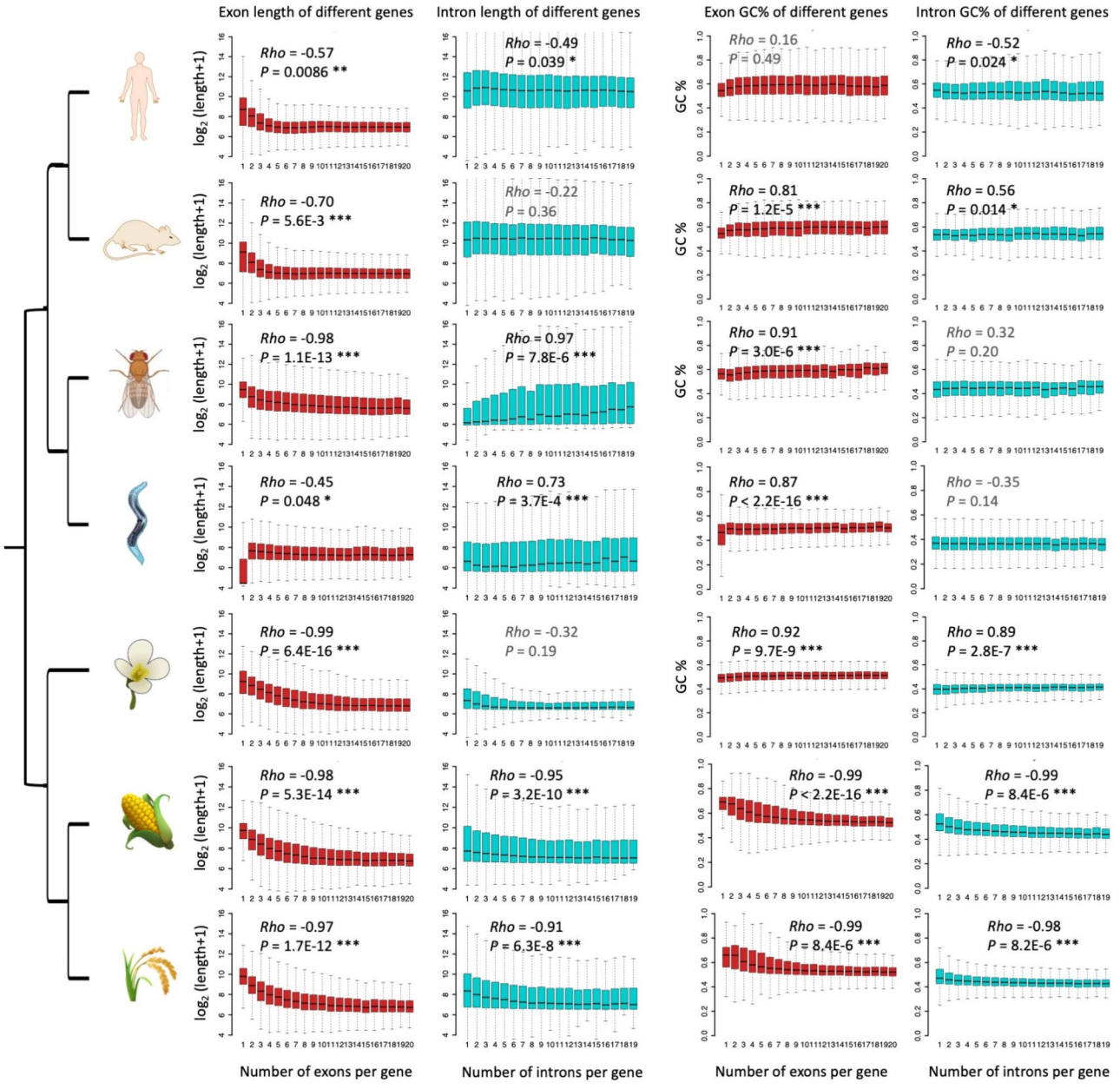

Fig. 2 Lengths and GC contents of exons and introns. The seven species are displayed in a phylogenetic order with unscaled branch length. Exon and intron lengths are shown with a $\log_2$(length + 1) transformation. The X-axis is the number of exons/introns per gene. That means, the exon/intron lengths (Y-axis) are grouped within these genes (each group is a box) and displayed separately in the boxplot. Spearman's correlation is calculated for each plot. The statistics of insignificant ones are colored in grey. Only genes with no more than 20 exons (19 introns) are shown. The reason for this is provided in the main text

the representative results of genes with two, six, ten, and 20 exons. We first focus on the length of exons. In all seven species, regardless of how many exons a gene has, the median length of last exons is always longer than those of all other exons (Fig. 3). Then, in six of the seven species (except worm), the first exon seems to be slightly longer than internal exons, and those internal exons do not exhibit differential length with each other (Fig. 3).

For last exons, their "length difference" over other exons increases with exon number per gene. For example, for genes with two exons, the last (means the second) exon does not show amazingly longer lengths than the first exon (Fig. 3, first column). But for genes with 20 exons, last exons are almost 5 ~ 10 times longer than the internal exons (Fig. 3, see the last column, note that the Y-axis is $\log_2$ transformed). Moreover, although the overall exon length decreases with exon number (Fig. 2),

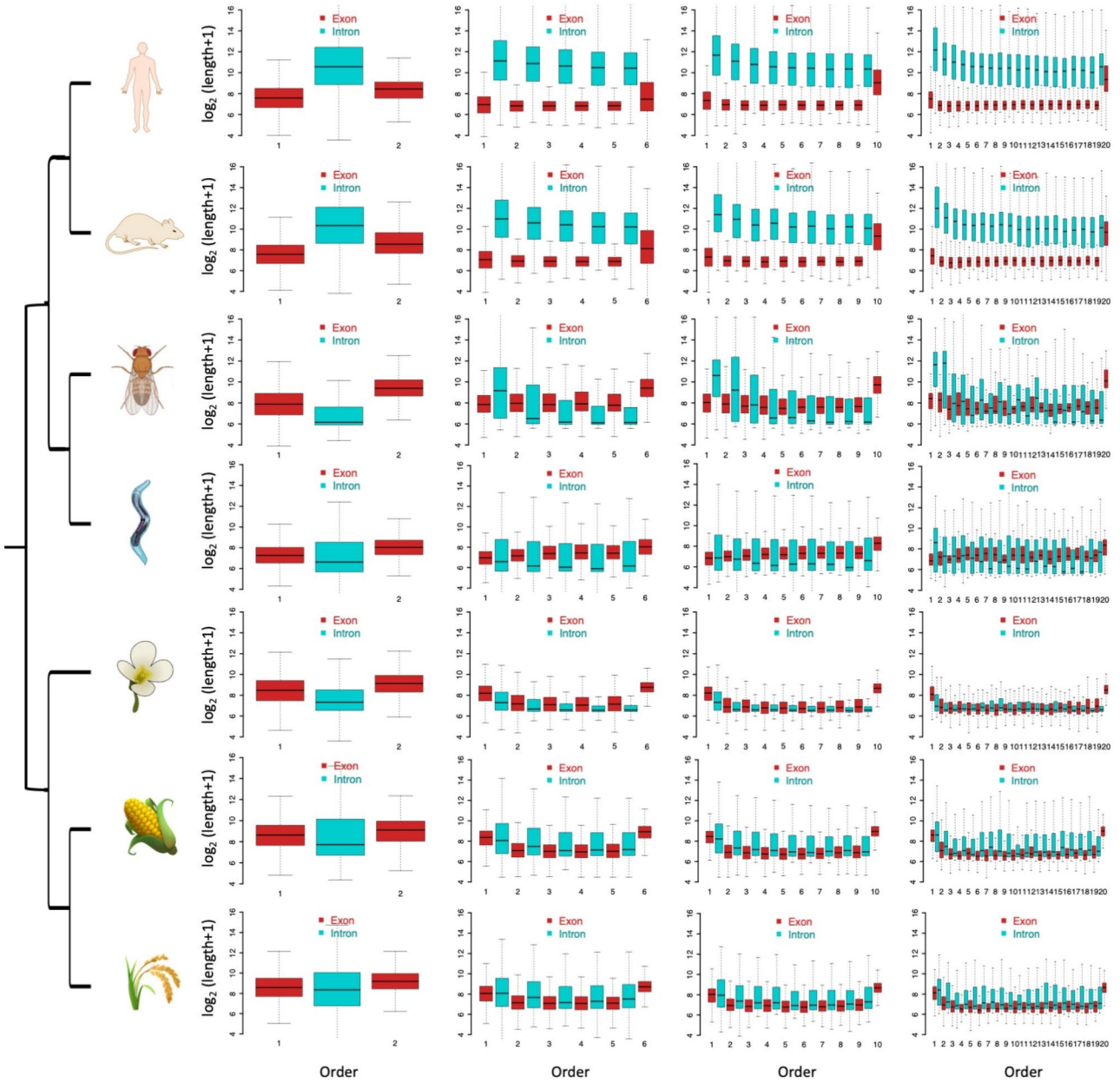

Fig. 3 Lengths of exons and introns in order. The seven species are displayed in a phylogenetic order with unscaled branch length. The four columns represent the genes with two, six, ten, and 20 exons, respectively. Exons and introns are displayed in order (from 5′ to 3′). Exons are in red and introns are in blue. Last exons are significantly longer than other exons in all cases, $P < 0.001$, Wilcoxon rank sum tests. The decrease of intron length is measured by Spearman correlation test. We obtained $P < 0.001$ in all cases with at least five introns

the length of last exons is actually increasing with exon length. That is to say, the 20th exon of genes with 20 exons is longer than the 10th exon of genes with 10 exons, which is further longer than the 6th exon of genes with 6 exons, and so on (see last exons in each panel of Fig. 3). This finding indicates a potential natural selection force favoring a longer length of last exons, and the strength of this selection force might increase with the number of exons per gene. The biological significance behind this phenomenon will be discussed in the next section.

Next, we looked at the intron length. Except worm, we found that the length of intron decreases from 5' to 3' regardless of how many introns a gene has (Fig. 3). This is a general confirmation of the "long first intron" phenomenon [24, 25, 34, 35], but we stress that *C. elegans* is an exception. Moreover, compared to the dramatic difference between the lengths of last exon and internal exons, the intron length seems to decrease mildly from 5' to 3' (Fig. 3). This suggests that intron length might

be governed by positional effect while the length of last exons might be constrained by its essential identity (3'UTR).

Canonical splicing motifs are avoided in 3'UTR, leading to the elongation of last exons

In all tested species, the median length of last exons is the longest regardless of how many exons a gene has (Fig. 3). We wonder what might be the evolutionary driving force triggering this pattern? In fact, since last exons mainly contain (or are included in) 3'UTR, we surmise that this property of long exon is related to the nature of 3'UTR. For example, in human and mouse, the intron-dependent nonsense-mediated decay (NMD) pathways will discourage the presence of introns (splicing junctions) within the 3'UTR [36]. However, this theory does not explain the same pattern observed in fruitfly and thale cress where the intron-dependent NMD mechanism is absent [36]. Moreover, although last exons are long, we still need a proper control to quantitatively demonstrate the avoidance of splicing junctions in 3'UTR.

We first noticed that the length of 3'UTR is significantly longer than the length of 5'UTR (Fig. 4A), agreeing with our common notion. However, the differential length between 3'UTR and 5'UTR is less striking compared to the difference between last exons and the other exons, or between last exons and first exons (Fig. 3). This raises a possibility that the elongation of last exons might enable them to contain the entire 3'UTR, and then the 3'UTR will be free from splicing junctions. 5'UTR can be used as a control when testing this hypothesis.

To clarify the relationships between UTRs and exons and test our hypothesis on the splicing-avoidance in 3'UTR, we retrieved the genes with at least three exons. We classified the genes containing 5'UTR into two classes (Fig. 4B). Case1, single-exon 5'UTR, meaning that the 5'UTR is contained in (or equal to) the first exon; case2, multi-exon 5'UTR, meaning that 5'UTR extends beyond the first exon. The same criteria were applied to classification of the genes with 3'UTR (Fig. 4B). Among the genes with at least three exons in human genome, 16,305 genes have 5'UTR and 7,341 (45.0%) of them have multi-exon 5'UTR, belonging to case2. In sharp contrast, for genes with 3'UTR, this fraction is only 2727/16,497 = 16.5% (Fig. 4A). Similarly, in other species, 3'UTR has significantly lower fraction to be multi-exon compared to 5'UTR (Fig. 4A). This pattern indicates that splicing junctions (or splicing events) are significantly avoided in 3'UTR.

In addition to the annotated splicing junctions, we also looked for canonical splicing motifs in the UTRs. Canonical introns typically have GU-AG motif at both ends (Fig. 5A). We counted the number of canonical splicing motifs in 5'UTR and 3'UTR, and used the number of motifs per Kb to represent motif density. In all species, we found that the canonical motif density is lower in 3'UTR compared to 5'UTR (Fig. 5B). This again suggests a strong avoidance for even the potential splicing sites in 3'UTR.

To understand why splicing is not welcome in 3'UTR, we consider two major functions of 3'UTR. (1) To be subjected to prevalent N^6-methyladenosine (m^6A) modification [37]; and (2) To be targeted by microRNAs [38]. In both animals and plants, m^6A modification sites are strongly enriched in 3'UTR [39, 40]. Although the methylation of adenosines mediated by the methyltransferases takes place in the nucleus [37, 41, 42], the functions of m^6A, such as regulating mRNA stability [43] and translation efficiency [42, 44], are exerted in cytoplasm. The 3'UTR m^6A sites are crucial to the mRNA regulation [39, 44]. It is possible that the suppression of splicing junctions in 3'UTR is driven by the avoidance of spliceosome-methyltransferase interaction. This interaction will either affect splicing or reduce m^6A modification in 3'UTR, disrupting the cellular system. In contrast, the microRNA-3'UTR interaction takes place post splicing, mainly in cytoplasm. It is unlikely that microRNA can constrain the distribution of splicing junctions in 3'UTR. Taken together, the m^6A hypothesis explains the underrepresentation of multi-exon 3'UTR, presumably by purifying selection on intron insertions into 3'UTR. Under this evolutionary scenario, the extraordinarily long length of last exons is nicely explained.

Last exons have lower GC content

GC content is usually connected to codon usage bias, gene expression, mRNA translation, RNA structure, and gene conversion [45–47]. Overall, in human and mouse, exons have comparable GC content to introns no matter how many exons/introns a gene has; but in fruitfly, worm, and plants, exons have remarkably higher GC content than introns (Fig. 2). This might be due to the selection on codon usage bias and mRNA translation. Fruitfly has much larger effective population size (Ne) = 1,150,000 [48] than human (20,974) [48] and mouse (160,000) [49], which leads to stronger selection on codon usage bias in fruitfly [50]. Besides, although the *Ne* of thale cress is variable across different populations [51], the selection on codon usage bias is still seen in this model plant [52].

Next, we displayed the GC content of exons/introns from 5' to 3' (Fig. 6 for the representative results of genes with two, six, ten, and 20 exons). For convenience, we divided the exons into first exons, last exons, and internal exons. The genes with only two exons will be discussed separately. We found two shared patterns among seven species: (1) last exons have lower median GC content than all internal exons; (2) the internal exons do not show remarkable difference with each other (Fig. 6). Then,

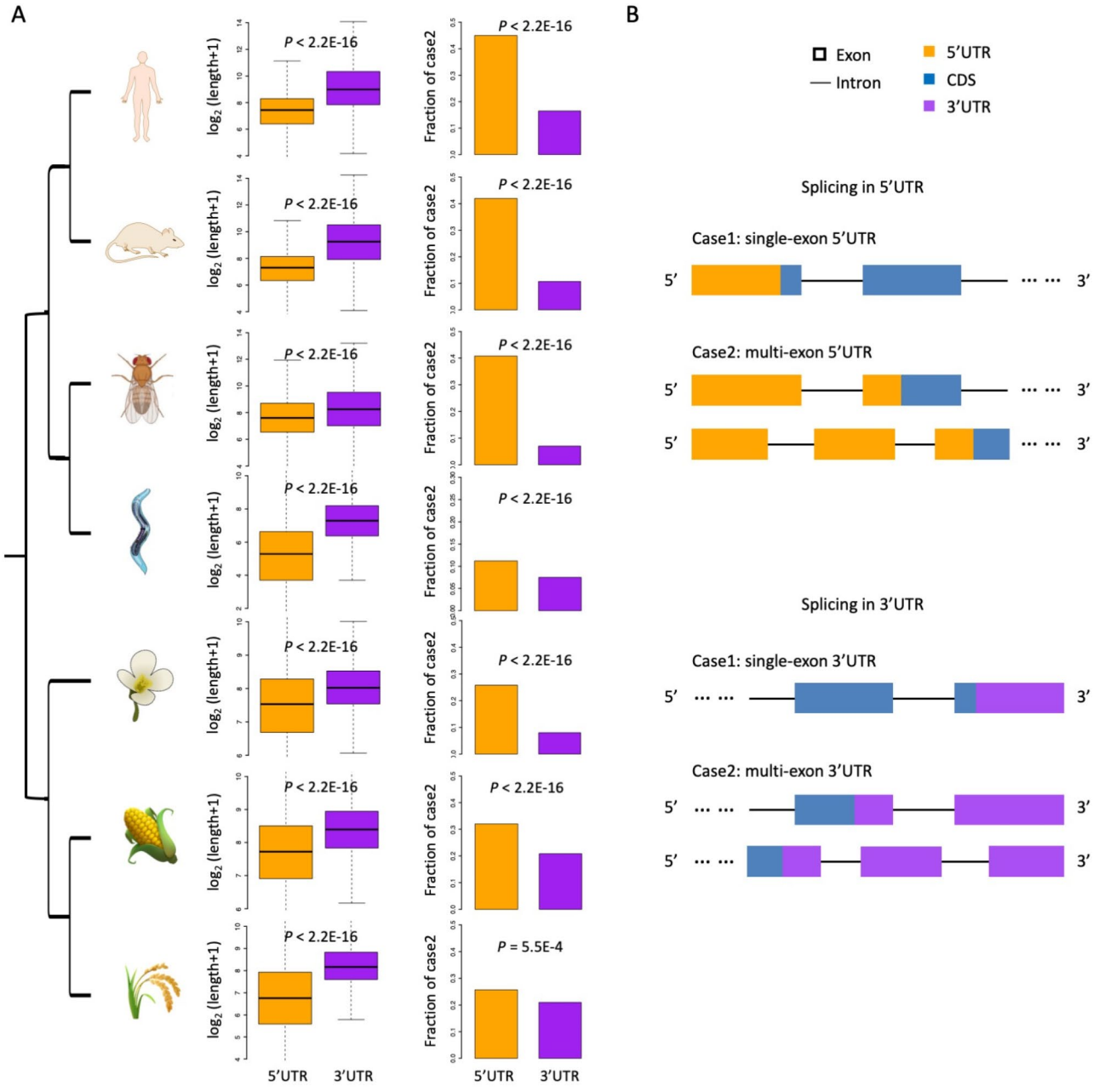

Fig. 4 The relationship between UTRs and exons. (**A**) Boxplots show the comparison between length distributions of 5'UTR and 3'UTR. *P* values were obtained by Wilcoxon rank sum tests. Barplots show the fraction of case2: multi-exon UTRs. *P* values were obtained by Fisher's exact tests. (**B**) Schematic diagram showing the definition of case1 and case2. In case1, UTR is contained in a single exon. In case2, UTR covers more than one exon

some species-specific patterns were also found: in mammals, corn, and rice, the first exon has higher GC content than internal exons; in fruitfly, the first exon has lower GC content than internal exons; and in thale cress and worm, the first exon has comparable GC content with internal exons (Fig. 6).

Although the GC content exhibits a correlation with the position of the exons, we found no evidence for such correlation between intronic GC content and the order of introns (Fig. 6). This pattern is plausible. As we have stated, the exonic GC content is related to codon usage bias, gene expression, and mRNA translation [45–47], but the intronic GC content has less positional effect. The effective population size is larger for fruitfly compared to mammals [48, 49] and thus natural selection on *cis* features determining codon bias and translation might also be efficient. Most eukaryotes including animals and plants enrich G/C-ending synonymous codons, and the high GC content in CDS (internal exons) is favorable for efficient decoding and fast translation elongation [28, 53].

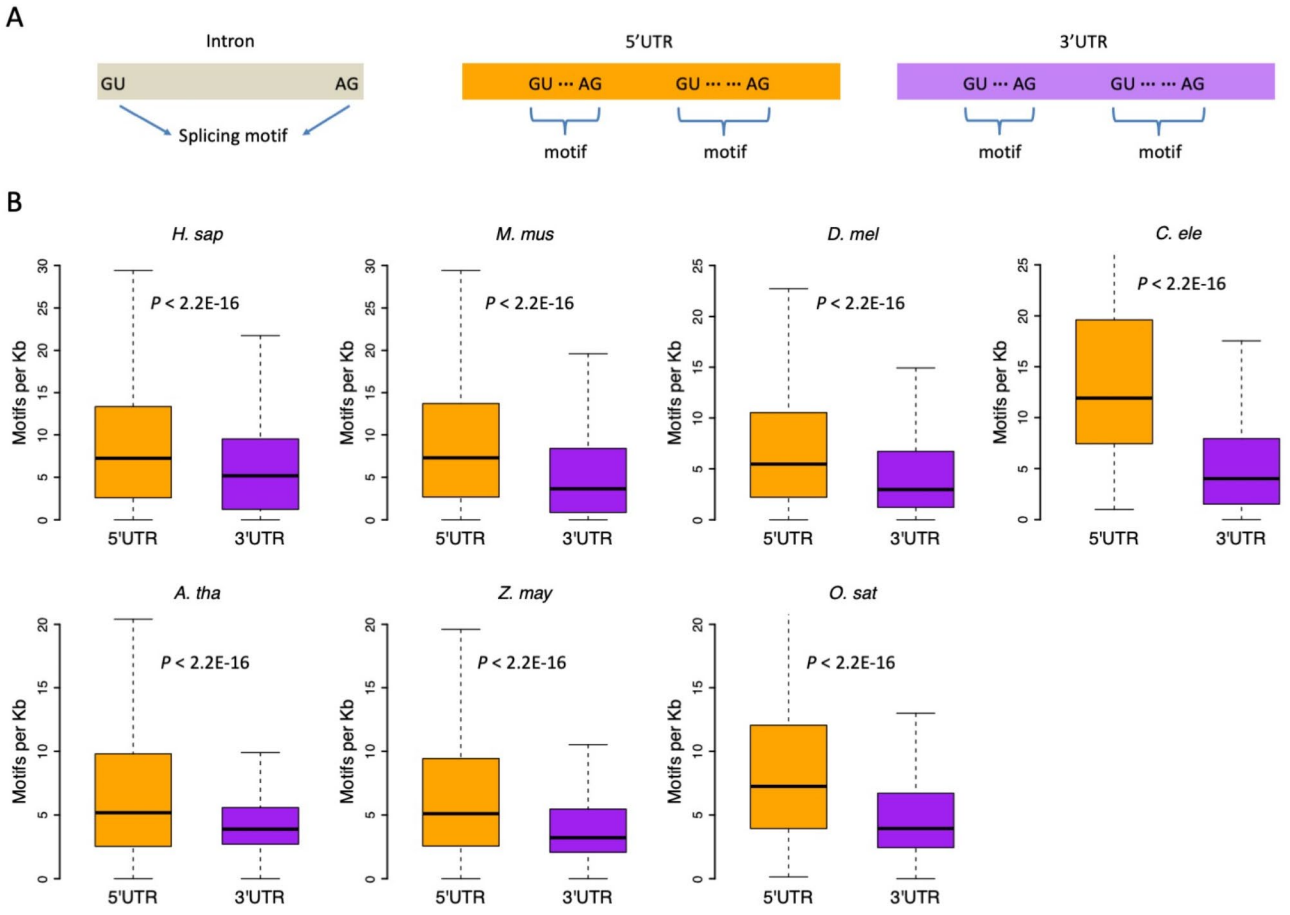

Fig. 5 Canonical splicing motif is underrepresented in 3'UTR. (**A**) Scheme for calculating the number of canonical splicing motifs in UTR. (**B**) Boxplots comparing the number of splicing motifs per Kb. *P* values were obtained by Wilcoxon rank sum tests

For the first exon, representing 5'UTR, a low GC content will unwind the RNA secondary structure, facilitating the scanning ribosome to find the start codon. Therefore, low GC content in 5'UTR is usually correlated with high translation initiation efficiency [54]. 5'UTR and CDS have opposite preference on GC content to achieve high initiation or elongation efficiency. In fruitfly, GC content is high for internal exons and low for the first exon (Fig. 6), supporting the selection on high translation efficiency [48]. In contrast, in human and mouse where *Ne* is smaller, the overall GC content in internal exons is only slightly higher than the intronic GC, and that the first exon (5'UTR) even has high GC content which is unfavorable for efficient translation initiation (Fig. 6).

Discussion

In this work, we relied on seven well-annotated reference genomes of animals and plants, and made three major findings: (1) Canonical splicing events/motifs are strongly underrepresented in 3'UTR. (2) Last exons have lower GC content. (3) In *D. melanogaster* rather than other species, the first exon has lower GC content than internal exons.

Then, although the long first intron and long last exon have already been reported in a few species [24–26], here we raise several points that need to be added to the established knowledges. The decrease of intron length is mild from 5' to 3', while last exons show a dramatic increase in length. This suggests that while intron length is position-dependent, exon length is related to the identity of a particular exon. For different introns in the same gene, the 5' introns are closer to the promoter of a gene and thus they can contain more regulatory elements to regulate the transcription [27]. This tendency is not necessarily restricted to the first intron. In fact, as long as a regulatory element works, it can be located in any introns. The position effect lets natural selection favor the regulatory elements to 5' introns. In sharp contrast, since exon length does not show a gradual positional change, it is likely that the constraint on the long last exon is only related to the nature (identity) of last exons, which is, 3'UTR. The steric effect between spliceosome and methylation machineries is a possible constraint.

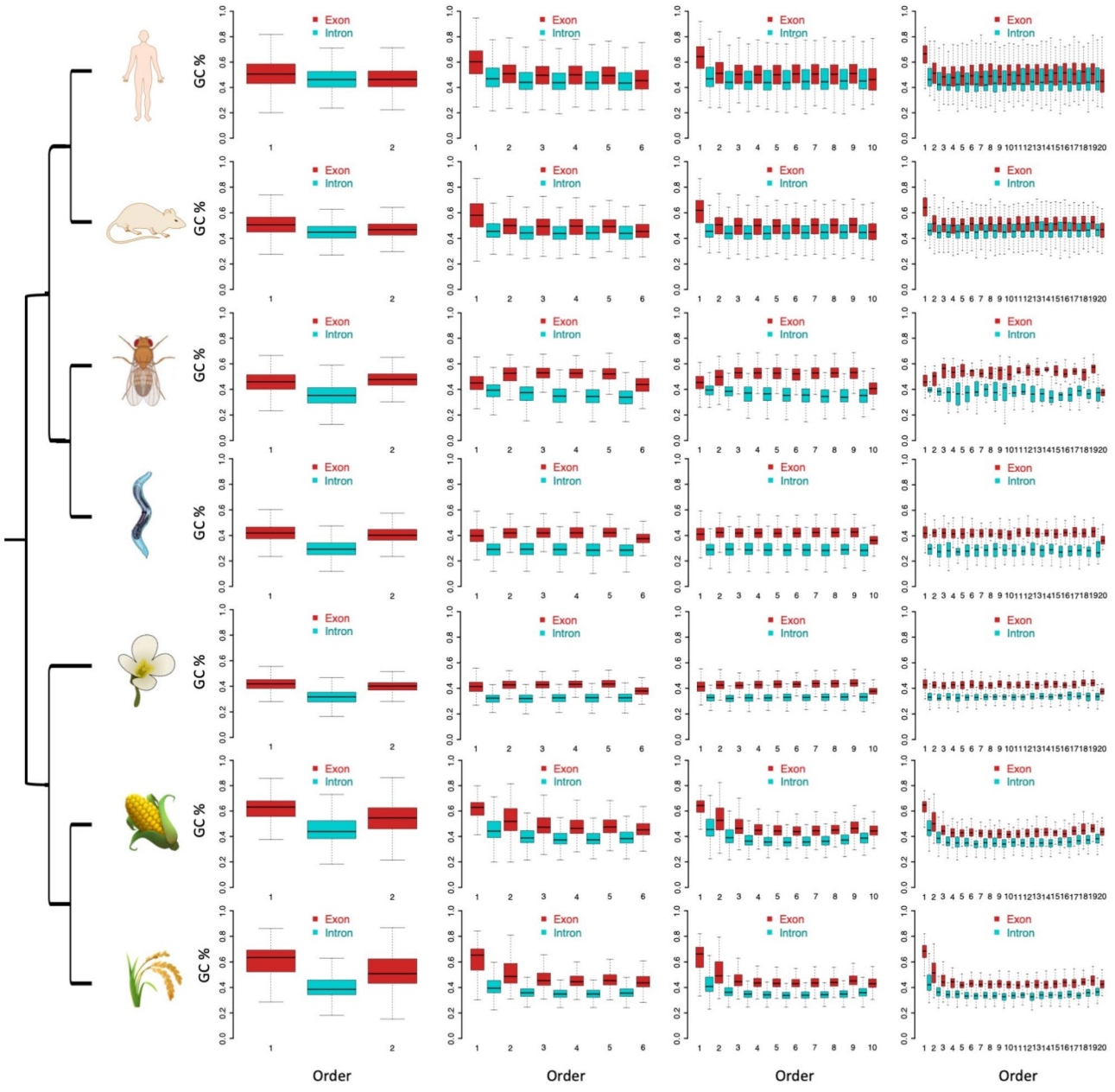

Fig. 6 GC content of exons and introns in order. The seven species are displayed in a phylogenetic order with unscaled branch length. The four columns represent the genes with two, six, ten, and 20 exons, respectively. Exons and introns are displayed in order (from 5′ to 3′). Exons are in red and introns are in blue. The differential GC content between exons were measured by Wilcoxon rank sum tests, and $P < 0.05$ was regarded as significant

One may consider that the underrepresentation of splicing junctions in 3'UTR can also be driven by the avoidance of spliceosome interacting with the RNA editing enzymes like ADARs [55, 56]. The ADAR-mediated A-to-I RNA editing is also a highly abundant RNA modification [57–59]. However, different from the enrichment of m^6A in 3'UTRs, A-to-I RNA editing sites have species-specific preferential locations. Mammalian RNA editing sites in studied species like humans [60, 61], mice [62], and pigs [63] are majorly located in repetitive regions and introns, while in insects (like true bugs [64] and bees [65]) and cephalopods (like squids [66] and octopuses [67]), RNA editing is enriched in coding sequence. None of the known species show a striking enrichment of A-to-I editing in 3'UTR. Moreover, plants mainly have C-to-U RNA editing in chloroplast and mitochondrial genes which do not involve the 3'UTR regulation at all [68, 69]. Thus, it is unlikely that RNA editing is a main force restricting the splicing in 3'UTR.

Last but not least, recent breaking news report that in bacteria, the *de novo* coding sequences can be synthesized from reverse transcription upon virus infection [70,

71]. This indicates that the reference genome of a species is far more complex than a linear sequence. More hidden information can be unraveled not only by the serendipitous experimental evidence, but also by careful and systematic scrutinization of the genome sequence.

conclusions

In seven representative and well-annotated reference genomes, canonical splicing events/motifs are strongly underrepresented in 3'UTR. Last exons have lower GC content. Comparing with other species, first exons in *D. melanogaster* genes demonstrate lower GC content than internal exons.

Materials and methods

Data acquisition

The reference genomes (fasta format) and the matched annotation files (gtf/gff format) were downloaded from the following addresses. *Homo sapiens*: Ensembl database genome version hg38 GRCh38.85 (https://asia.ensembl.org/, link to the data https://ftp.ensembl.org/pub/release-85/fasta/homo_sapiens/). *Mus musculus*: Ensembl database genome version mm10 GRCm38.85 (https://asia.ensembl.org/, link to the data https://ftp.ensembl.org/pub/release-85/fasta/mus_musculus/). *Drosophila melanogaster*: FlyBase genome version r6.06 (https://flybase.org/, link to the data: https://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.06_FB2015_03/). *Caenorhabditis elegans*: Ensembl database, genome version WBcel235 (https://ftp.ensembl.org/pub/release-85/fasta/caenorhabditis_elegans/). *Arabidopsis thaliana*: Ensembl plants genome version TAIR10.59 (https://ftp.ensemblgenomes.ebi.ac.uk/pub/current/plants/fasta/arabidopsis_thaliana/dna/). *Zea mays* genome version Zm-B73-REFERENCE-NAM-5.0 (https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-60/fasta/zea_mays/) and *Oryza sativa* genome version IRGSP-1.0 (https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-60/fasta/oryza_sativa/) were also downloaded from Ensembl plants. Note that the *Drosophila* genome/annotation downloaded from Ensembl (version BDGP6) is essentially retrieved from FlyBase. Code for data processing and analysis is available in Additional file 2: Supplementary Data S1.

Processing the genome annotation file

To calculate the lengths of different exons and introns, we first need to know the genomic coordinate of the intervals of each exon/intron. The annotation file in gtf or gff format is 1-based, meaning that the length should be calculated as end − start + 1. We manually convert gtf/gff to bed format where the start position is 0-based so that the length of the interval will be end − start. Each interval in the annotation file is labeled with the feature (gene/transcript/exon/CDS/UTR). In 0-based format, the subtraction to calculate length was done by awk 'length = $3-$2'. Intron region is not provided in the annotation file but we can infer its coordinates from the interval of exons. For detail, please refer to the code (Additional file 2: Supplementary Data S1).

Number of transcripts per gene and number of exons per transcript

To know how many transcripts does a gene have, the most direct way is to search for a gene name and see how many different transcripts were annotated. Our idea is basically the same, but we did it in a bioinformatic manner. The genome annotation file (gft/gff) contains the transcript IDs and gene IDs. We deduplicate (or unique, as a verb) the transcript IDs, making each transcript ID a line. Then we can see that some different transcript IDs have an identical gene ID (e.g. geneA). We count the appearance of this geneA, the outcome is the number of transcripts geneA has. This example shows how to obtain the number of transcripts for geneA. For all genes in the genome, we use Linux command "uniq -c" or equivalently, "table" in R language, to count the genome-wide results of transcripts per gene. The same goes for the number of exons per transcript. We only need to make each unique exon ID per line and count the appearance of each transcript ID. For detail please see the code provided in Additional file 2: Supplementary Data S1s>.

GC content and searching of canonical splicing motifs

The above analyses only require the position and number information and do not require the sequence information: because one could envision that even without knowing the sequence, we could still know how many transcripts a gene has and how many exons a transcript has. But to calculate the GC content of a genomic region, knowing its genomic coordinate (position) is not enough. The sequence information is in the reference genome file in fasta format. Bedtools v2.28.0 [72] was used to process the sequence file for a given interval. "Bedtools getfasta" extracts the sequence of an interval assigned in the bed file, and "Bedtools nuc" directly reports the numbers of each nucleotide of an assigned interval (see code Additional file 2: Supplementary Data S1). Then, the GC content can be calculated based on the numbers of each nucleotide. The searching of GU-AG splicing motifs in a given sequence was accomplished by "str_count" function of "stringr" package in R language. We understand that the definition of splicing motifs usually requires an upstream branch site A in addition to the canonical GU-AG motif [73]. Here, we first realized the difficulty in searching for the branch site A due to that its distance to the splicing site is variable. In fact, the probability of see an adenosine within an upstream (e.g. 0 ~ 50 bp) region is extremely high, then we believe that adding this criterion

will not discard many of our candidates. Second, we logically argue that the more stringent pipeline has already been applied to annotate the splicing sites in the reference genomes. Then, repeating the same stringent pipeline will not produce more splicing sites. Since our purpose is to find the "relic" of splicing motifs in UTRs that possibly has already been disabled by natural selection, searching for canonical GU-AG motif might be enough.

Statistics and graphical works

Statistical tests and graphical works were all accomplished in R studio (R version 3.6.3). Code used in this study is available in Additional file 2: Supplementary Data S1.

Abbreviations

A-to-I	Adenosine-to-inosine
C-to-U	Cytidine-to-uridine
CDS	Coding sequence
m^6A	N^6-methyladenosine
MFE	Minimum free energy
Ne	Effective population size
UTR	Untranslated region

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11504-1.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

The computational work is supported by High-performance Computing Platform of China Agricultural University. We thank the platform for the computational support. We thank the lab members for their suggestions to this work.

Author contributions

Conceptualization & supervision: Y.D. and Q.C.Data analysis: Y.D. and Q.C.Writing – original draft: Y.D. and Q.C.Writing – review & editing: Y.D. and Q.C.All authors approved the submission of this manuscript.

Funding

This study is financially supported by Beijing Natural Science Foundation (Natural Science Foundation of Beijing Municipality) no.5254030, China Postdoctoral Science Foundation 2024M760144, and the 2115 Talent Development Program of China Agricultural University.

Data availability

The reference genomes (fasta format) and the matched annotation files (gtf/gff format) were downloaded from the following addresses. ***Homo sapiens***: Ensembl database genome version hg38 GRCh38.85 (https://asia.ensembl.org/, link to the data https://ftp.ensembl.org/pub/release-85/fasta/homo_sapiens/). ***Mus musculus***: Ensembl database genome version mm10 GRCm38.85 (https://asia.ensembl.org/, link to the data https://ftp.ensembl.org/pub/release-85/fasta/mus_musculus/). ***Drosophila melanogaster***: FlyBase genome version r6.06 (https://flybase.org/, link to the data: https://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.06_FB2015_03/). ***Caenorhabditis elegans***: Ensembl database, genome version WBcel235 (https://ftp.ensembl.org/pub/release-85/fasta/caenorhabditis_elegans/). ***Arabidopsis thaliana***: Ensembl plants genome version TAIR10.59 (https://ftp.ensemblgenomes.ebi.ac.uk/pub/current/plants/fasta/arabidopsis_thaliana/dna/). ***Zea mays*** genome version Zm-B73-REFERENCE-NAM-5.0 (https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-60/fasta/zea_mays/) and ***Oryza sativa*** genome version IRGSP-1.0 (https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-60/fasta/oryza_sativa/) were also downloaded from Ensembl plants. Note that the ***Drosophila*** genome/annotation downloaded from Ensembl (version BDGP6) is essentially retrieved from FlyBase. Code for data processing and analysis is available in Additional file 2: Supplementary Data S1.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 2 January 2025 / Accepted: 19 March 2025
Published online: 24 March 2025

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.
2. Arabidopsis Genome I. Analysis of the genome sequence of the flowering plant *Arabidopsis Thaliana*. Nature. 2000;408(6814):796–815.
3. Mouse Genome Sequencing C, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420(6915):520–62.
4. Massimino M, Martorana F, Stella S, Vitale SR, Tomarchio C, Manzella L, Vigneri P. Single-cell analysis in the omics era: technologies and applications in cancer. Genes (Basel). 2023;14(7):1330.
5. Di Girolamo F, Lante I, Muraca M, Putignani L. The role of mass spectrometry in the omics era. Curr Org Chem. 2013;17(23):2891–905.
6. Xue R, Zhang Q, Cao Q, Kong R, Xiang X, Liu H, Feng M, Wang F, Cheng J, Li Z, et al. Liver tumour immune microenvironment subtypes and neutrophil heterogeneity. Nature. 2022;612(7938):141–7.
7. Eisenberg E. Bioinformatic approaches for identification of A-to-I editing sites. Curr Top Microbiol Immunol. 2012;353:145–62.
8. Liu Z, Quinones-Valdez G, Fu T, Huang E, Choudhury M, Reese F, Mortazavi A, Xiao X. L-GIREMI uncovers RNA editing sites in long-read RNA-seq. Genome Biol. 2023;24(1):171.
9. Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, Li JB. Identifying RNA editing sites using RNA sequencing data alone. Nat Methods. 2013;10(2):128–32.
10. Zhang Y, Duan Y. Genome-wide analysis on driver and passenger RNA editing sites suggests an underestimation of adaptive signals in insects. Genes (Basel). 2023;14(10):1951.
11. Duan Y, Xu Y, Song F, Tian L, Cai W, Li H. Differential adaptive RNA editing signals between insects and plants revealed by a new measurement termed haplotype diversity. Biol Direct. 2023;18(1):47.
12. Xu Y, Liu J, Zhao T, Song F, Tian L, Cai W, Li H, Duan Y. Identification and interpretation of A-to-I RNA editing events in insect transcriptomes. Int J Mol Sci. 2023;24(24):17126.
13. Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. Cell. 2014;156(6):1324–35.
14. Wei L. Selection on synonymous mutations revealed by 1135 genomes of Arabidopsis Thaliana. Evol Bioinform Online. 2020;16:1176934320916794.
15. Duan Y, Ma L, Zhao T, Liu J, Zheng C, Song F, Tian L, Cai W, Li H. Conserved A-to-I RNA editing with non-conserved recoding expands the candidates of functional editing sites. Fly (Austin). 2024;18(1):2367359.
16. Ma L, Zheng C, Liu J, Song F, Tian L, Cai W, Li H, Duan Y. Learning from the codon table: convergent recoding provides novel Understanding on the evolution of A-to-I RNA editing. J Mol Evol 2024;92.

17. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. The genome sequence of *Drosophila melanogaster*. Science. 2000;287(5461):2185–95.
18. Korstanje C. The human genome project: Understanding the role of inflammation in disease and disease prevention. Curr Opin Investig Drugs. 2006;7(11):964–5.
19. Babenko VN, Chadaeva IV, Orlov YL. Genomic landscape of CpG rich elements in human. BMC Evol Biol. 2017;17(19). https://doi.org/10.1186/s12862-12016-10864-12860.
20. Rogozin IB, Sverdlov AV, Babenko VN, Koonin EV. Analysis of evolution of exon-intron structure of eukaryotic genes. Brief Bioinform. 2005;6(2):118–34.
21. Dvorak P, Hanicinec V, Soucek P. The position of the longest intron is related to biological functions in some human genes. Front Genet. 2022;13:1085139.
22. Glick L, Castiglione S, Loewenthal G, Raia P, Pupko T, Mayrose I. Phylogenetic analysis of 590 species reveals distinct evolutionary patterns of intron-exon gene structures across eukaryotic lineages. Mol Biol Evol. 2024;41(12):msae248.
23. Sakharkar MK, Chow VT, Kangueane P. Distributions of exons and introns in the human genome. Silico Biol. 2004;4(4):387–93.
24. Zhu L, Zhang Y, Zhang W, Yang S, Chen JQ, Tian D. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genomics. 2009;10:47.
25. Bradnam KR, Korf I. Longer first introns are a general property of eukaryotic gene structure. PLoS ONE. 2008;3(8):e3093.
26. Kalari KR, Casavant M, Bair TB, Keen HL, Comeron JM, Casavant TL, Scheetz TE. First exons and introns–a survey of GC content and gene structure in the human genome. Silico Biol. 2006;6(3):237–42.
27. Duret L. Why do genes have introns? Recombination might add a new piece to the puzzle. Trends Genet. 2001;17(4):172–5.
28. Chu D, Wei L. Trade-off between cost and efficiency during mRNA translation is largely driven by natural selection in angiosperms. Plant Syst Evol. 2020;306(6):92.
29. Duan Y, Ma L, Song F, Tian L, Cai W, Li H. Autorecoding A-to-I RNA editing sites in the *Adar* gene underwent compensatory gains and losses in major insect clades. RNA. 2023;29(10):1509–19.
30. Ma L, Zheng C, Xu S, Xu Y, Song F, Tian L, Cai W, Li H, Duan Y. A full repertoire of hemiptera genomes reveals a multi-step evolutionary trajectory of auto-RNA editing site in insect *Adar* gene. RNA Biol. 2023;20(1):703–14.
31. Ohno S. Ancient linkage groups and frozen accidents. Nature. 1973;244(5414):259–62.
32. Rodriguez-Trelles F, Tarrio R, Ayala FJ. Origins and evolution of spliceosomal introns. Annu Rev Genet. 2006;40:47–76.
33. Roy SW, Gilbert W. The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet. 2006;7(3):211–21.
34. Bernards A, Rubin CM, Westbrook CA, Paskind M, Baltimore D. The first intron in the human *c-abl* gene is at least 200 kilobases long and is a target for translocations in chronic myelogenous leukemia. Mol Cell Biol. 1987;7(9):3231–6.
35. Robinson-Thiewes S, Kimble J. *C. elegans mpk-1b* long first intron enhances MPK-1B protein expression. *MicroPubl Biol* 2021, 2021.
36. Hong X, Scofield DG, Lynch M. Intron size, abundance, and distribution within untranslated regions of genes. Mol Biol Evol. 2006;23(12):2392–404.
37. Zhao BS, Nachtergaele S, Roundtree IA, He C. Our views of dynamic N(6)-methyladenosine RNA methylation. *RNA* 2018, 24(3):268–272.
38. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136(2):215–33.
39. Wang G, Li H, Ye C, He K, Liu S, Jiang B, Ge R, Gao B, Wei J, Zhao Y, et al. Quantitative profiling of m(6)A at single base resolution across the life cycle of rice and *Arabidopsis*. Nat Commun. 2024;15(1):4881.
40. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. Cell. 2012;149(7):1635–46.
41. Liu N, Dai Q, Zheng G, He C, Parisien M, Pan T. N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. Nature. 2015;518(7540):560–4.
42. Zhou J, Wan J, Gao X, Zhang X, Jaffrey SR, Qian SB. Dynamic m(6)A mRNA methylation directs translational control of heat shock response. Nature. 2015;526(7574):591–4.
43. Wang X, Lu Z, Gomez A, Hon GC, Yue Y, Han D, Fu Y, Parisien M, Dai Q, Jia G, et al. N6-methyladenosine-dependent regulation of messenger RNA stability. Nature. 2014;505(7481):117–20.
44. Wang X, Zhao BS, Roundtree IA, Lu Z, Han D, Ma H, Weng X, Chen K, Shi H, He C. N(6)-methyladenosine modulates messenger RNA translation efficiency. Cell. 2015;161(6):1388–99.
45. Chu D, Wei L. Characterizing the heat response of *Arabidopsis Thaliana* from the perspective of codon usage bias and translational regulation. J Plant Physiol. 2019;240:153012.
46. Harrison RJ, Charlesworth B. Biased gene conversion affects patterns of codon usage and amino acid usage in the *Saccharomyces sensu* stricto group of yeasts. Mol Biol Evol. 2011;28(1):117–29.
47. Zhou Z, Dang Y, Zhou M, Li L, Yu CH, Fu J, Chen S, Liu Y. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proc Natl Acad Sci U S A. 2016;113(41):E6117–25.
48. Gossmann TI, Keightley PD, Eyre-Walker A. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol Evol. 2012;4(5):658–67.
49. Salcedo T, Geraldes A, Nachman MW. Nucleotide variation in wild and inbred mice. Genetics. 2007;177(4):2277–91.
50. Campos JL, Zeng K, Parker DJ, Charlesworth B, Haddrill PR. Codon usage bias and effective population sizes on the X chromosome *versus* the autosomes in *Drosophila melanogaster*. Mol Biol Evol. 2013;30(4):811–23.
51. Gomaa NH, Montesinos-Navarro A, Alonso-Blanco C, Pico FX. Temporal variation in genetic diversity and effective population size of mediterranean and subalpine *Arabidopsis Thaliana* populations. Mol Ecol. 2011;20(17):3540–54.
52. Chu D, Wei L. Direct *in vivo* observation of the effect of codon usage bias on gene expression in *Arabidopsis* hybrids. J Plant Physiol. 2021;265:153490.
53. Hanson G, Coller J. Codon optimality, bias and usage in translation and mRNA decay. Nat Rev Mol Cell Biol. 2018;19(1):20–30.
54. Zhu L, Wang Q, Zhang W, Hu H, Xu K. Evidence for selection on SARS-CoV-2 RNA translation revealed by the evolutionary dynamics of mutations in UTRs and CDSs. RNA Biol. 2022;19(1):866–76.
55. Savva YA, Rieder LE, Reenan RA. The ADAR protein family. Genome Biol. 2012;13(12):252.
56. Xie Q, Duan Y. An ultimate question for functional A-to-I mRNA editing: why not a genomic G? J Mol Evol. 2025. https://doi.org/10.1007/s00239-00025-10238-00238.
57. Duan Y, Li H, Cai W. Adaptation of A-to-I RNA editing in bacteria, fungi, and animals. Front Microbiol. 2023;14:1204080.
58. Eisenberg E, Levanon EY. A-to-I RNA editing - immune protector and transcriptome diversifier. Nat Rev Genet. 2018;19(8):473–90.
59. Zhang P, Zhu Y, Guo Q, Li J, Zhan X, Yu H, Xie N, Tan H, Lundholm N, Garcia-Cuetos L, et al. On the origin and evolution of RNA editing in metazoans. Cell Rep. 2023;42(2):112112.
60. Liddicoat BJ, Piskol R, Chalk AM, Ramaswami G, Higuchi M, Hartner JC, Li JB, Seeburg PH, Walkley CR. RNA editing by ADAR1 prevents MDA5 sensing of endogenous DsRNA as nonself. Science. 2015;349(6252):1115–20.
61. Tan MH, Li Q, Shanmugam R, Piskol R, Kohler J, Young AN, Liu KI, Zhang R, Ramaswami G, Ariyoshi K, et al. Dynamic landscape and regulation of RNA editing in mammals. Nature. 2017;550(7675):249–54.
62. Licht K, Kapoor U, Amman F, Picardi E, Martin D, Bajad P, Jantsch MF. A high resolution A-to-I editing map in the mouse identifies editing events controlled by pre-mRNA splicing. Genome Res. 2019;29(9):1453–63.
63. Adetula AA, Fan X, Zhang Y, Yao Y, Yan J, Chen M, Tang Y, Liu Y, Yi G, Li K, et al. Landscape of tissue-specific RNA editome provides insight into co-regulated and altered gene expression in pigs (*Sus-scrofa*). RNA Biol. 2021;18(sup1):439–50.
64. Duan Y, Ma L, Liu J, Liu X, Song F, Tian L, Cai W, Li H. The first A-to-I RNA editome of hemipteran species *Coridius chinensis* reveals overrepresented recoding and prevalent intron editing in early-diverging insects. Cell Mol Life Sci. 2024;81:136.
65. Liu J, Zhao T, Zheng C, Ma L, Song F, Tian L, Cai W, Li H, Duan Y. An orthology-based methodology as a complementary approach to retrieve evolutionarily conserved A-to-I RNA editing sites. RNA Biol. 2024;21(1):29–45.
66. Alon S, Garrett SC, Levanon EY, Olson S, Graveley BR, Rosenthal JJ, Eisenberg E. The majority of transcripts in the squid nervous system are extensively recoded by A-to-I RNA editing. eLife. 2015;4:e05198.
67. Liscovitch-Brauer N, Alon S, Porath HT, Elstein B, Unger R, Ziv T, Admon A, Levanon EY, Rosenthal JJC, Eisenberg E. Trade-off between transcriptome plasticity and genome evolution in cephalopods. Cell. 2017;169(2):191–e202111.
68. Chu D, Wei L. The Chloroplast and mitochondrial C-to-U RNA editing in *Arabidopsis Thaliana* shows signals of adaptation. Plant Direct. 2019;3(9):e00169.

69. Duan Y, Cai W, Li H. Chloroplast C-to-U RNA editing in vascular plants is adaptive due to its restorative effect: testing the restorative hypothesis. RNA. 2023;29(2):141–52.
70. Tang S, Conte V, Zhang DJ, Zedaveinyte R, Lampe GD, Wiegand T, Tang LC, Wang M, Walker MWG, George JT et al. *De Novo* gene synthesis by an antiviral reverse transcriptase. Science 2024;386:eadq0876.
71. Wilkinson ME, Li D, Gao A, Macrae RK, Zhang F. Phage-triggered reverse transcription assembles a toxic repetitive gene from a noncoding RNA. Science 2024;386:eadq3977.
72. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
73. Xue C, Zhang H, Lin Q, Fan R, Gao C. Manipulating mRNA splicing by base editing in plants. Sci China Life Sci. 2018;61(11):1293–300.

Publisher's note